URL https://journal.ntt.co.jp/article/26661
DOI https://doi.org/10.60249/24061201

NTTコミュニケーションズ
イノベーションセンター　テクノロジー部門　担当課長

岩瀬義昌　Yoshimasa Iwase

「tsuzumi」を特化型にチューニングしてお客さまに届ける

2022年11月に生成AI（人工知能）であるChatGPTが発表され，急速な広がりを見せる中，GPTをはじめとする大規模言語モデル（LLM：Large Language Models）も急速に進化し，大規模化しています．こうした動きの中，NTTは，「日本語対応」「小型・軽量」「柔軟なチューニング」「マルチモーダル」を特長とするLLM，「**tsuzumi**」を2023年11月1日に発表しました．「**tsuzumi**」の特長を活かして，業界やユースケースに特化した専門的なLLMとして，市場展開が期待されています．「**tsuzumi**」の専門分野特化型へのチューニングは，NTTコミュニケーションズ，NTTドコモ，NTTデータ等の事業会社で対応して，お客さまにサービス・ソリューションとして提供されていきます．NTTコミュニケーションズ イノベーションセンターの岩瀬義昌氏に，チューニング等による特化のプロセス，技術的課題，そして技術のコアな部分の基本スキルと情報発信の大切さについて伺いました．

「tsuzumi」の特化においては，ファインチューンとRAGの最適化がポイント

現在，手掛けている技術の概要をお聞かせいただけますか．

NTTコミュニケーションズ（NTT Com）イノベーションセンターのGenerative AIプロジェクトで，生成AI（人工知能）の開発，検証，評価に取り組んでいます．生成AIを構成する大規模言語モデル（LLM：Large Language Models）として，NTTグループには2023年11月1日に発表され，2024年3月25日に商用開始が発表された「tsuzumi」があります．「tsuzumi」はNTT人間情報研究所において研究開発・実用化され，それをNTT Com，NTTデータなどのNTTグループ各社でお客さまのソリューションや専門分野に特化するようチューニングを行うといった連携により，商用提供されます．

私たちのプロジェクトでは，「tsuzumi」を含む生成AIに関して，CoE（Center of Excellence）として事業部支援と，出島/研究開発という2つの側面から取り組んでいます．CoEとしては，NTT Com社内のGenerative AIタスクフォースが実際のお客さまのシステムを組み上げる際に，そのユースケースに一番適合するかたちで「tsuzumi」をはじめとするAIの精度を高める支援をしています．特に「tsuzumi」に関するノウハウは研究所の支援を受けながら私たちで検証を重ねて，それをベースに実際のお客さま案件やプロダクトへのLLMの組み込みに適用するというかたちで支援しています．また，LLMを組み込む際に必要となるファインチューンや，文書検索モデルが外部のデータベースを参照し回答元の情報を選択し，LLMがその検索結果を理解し文脈に沿った回答を生成するRAG（Retrieval Augmented Generation）の実装といった技術を事業部で実現できるようにするための技術支援・育成も行っています．

「tsuzumi」の実装はどのようなプロセスでなされるのでしょうか．

「tsuzumi」の場合，NTT人間情報研究所において，世の中のコーパスデータを大量に収集し，そのデータにおける重複，記載ミス，表記揺れなどを検出し，削除・修正を行うクレンジング，LLMの重みを決定するための「事前学習」を行います．それを私たちのプロジェクトでは必要に応じて，データ収集，クレンジング，そして事前学習において定義されたモデルの選定を経て学習させる「継続事前学習」を行います．それに対してほぼ同様なプロセスにより専門領域に特化したデータを学習させる「ファインチューン」を行い，クラウド等任意の基盤に展開（デプロイ）し，推論テストを経て，モデルが公開されます．そのモデルを実際のシステムに組み込むことで一連のシステムとして実装され，試験・評価の後にサービス開始となります（**図1**）．

出島/研究開発では，図1の網掛け部分に対応して①「tsuzumi」他LLMの検証・評価，②ファインチューンやRAGに適した「情

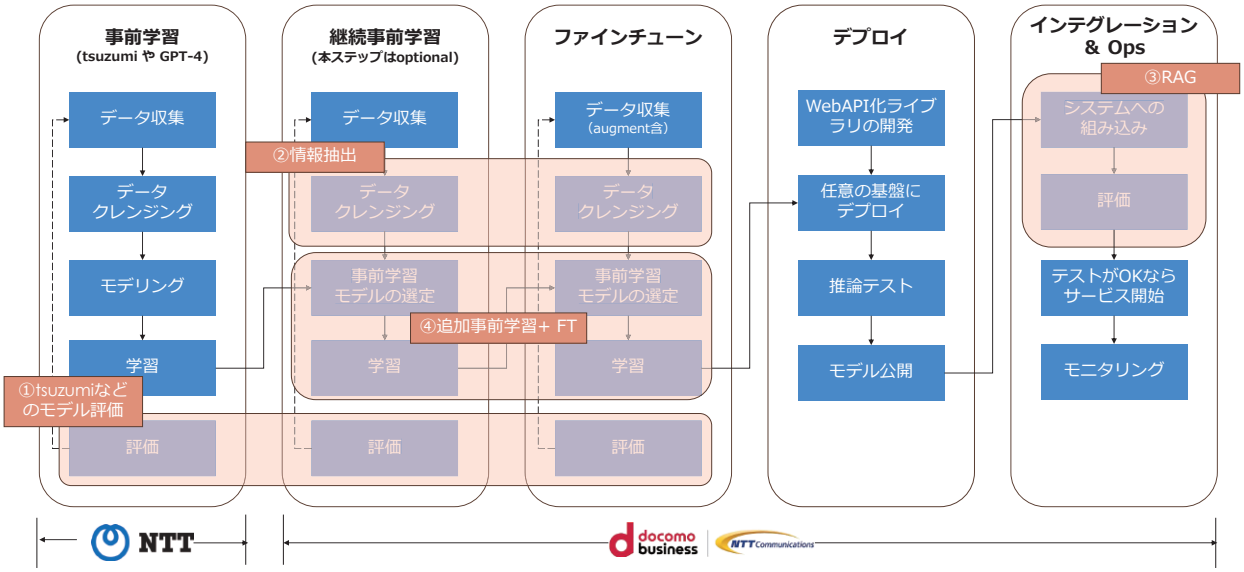

図1　「tsuzumi」の実装プロセスと技術開発テーマのマッピング

報抽出技術」の開発・評価，③高精度なRAG方式の開発・検証・評価（RAG Assessmentなど），④追加事前学習による特化モデル開発・評価の4つのテーマに取り組んでいます．

①「tsuzumi」他LLMの検証・評価については，新バージョンの「tsuzumi」がリリースされたときの検証・評価はもちろん行っていますが，「tsuzumi」のポジショニング確認と特長を踏まえた適用領域検討のために他のLLMについても行っています．評価は一例としてRakudaというベンチマークツールを使っています．ただRakudaには評価過程における癖があるため，他のベンチマークツール等も試行錯誤的に利用し，その結果も取り入れながら複合的に評価しています．さらに，さまざまなユースケースに組み込んで，そのときの精度・性能についての評価も行っています．

LLMは一般に，世の中のWebデータを使って学習しているので，例えば，社内や業界内の固有な知識は持っていません．そのため，社会実装に向けてこういった情報を使えるようにする必要があります．これは，学習するのであればファインチューン，そうでない場合はRAGにより行います．専門分野に特化したデータを学習させる際に，「tsuzumi」に対するファインチューンでは特化データを学習すると，すでに学習済の汎用的な知識が追い出されることがあります（破滅的忘却）．そのために，どのようなデータを用いて専門知識を学習させれば精度の高いLLMになるのかといった，試行錯誤が非常に重要な要素となります．また，RAGでは，データ形式に合わせて，LLMの推論に親和性の高い情報を適切に検索する必要があります．

さらに，LLMの学習・活用には高品質なデータが不可欠である一方で，画像などを多く含むドキュメント活用の困難さがあります．これらの課題に対する取り組みが，②ファインチューンやRAGに適した「情報抽出技術」の開発・評価です．

ファインチューンとRAGのどちらも独自知識の組み込みに効果的であることは説明したとおりですが，一方で，生成AIを実際のシステムに組み込む際には，RAGのほうが効果的という結果が多く出始めています．そこで，③高精度なRAG方式の開発・検証・評価（RAG Assessmentなど）の取り組みを行っています．

④追加事前学習による特化モデル開発・評価については，追加事前学習により，流通，金融といった業界の特化モデルを開発し，評価する取り組みです．業界特化モデルについては実験を始めており，お客さまによっては業界にとどまらず自社に特化したモデルに関するご要望をいただくこともあります．その際にお客さま固有データを活用して，ファインチューンによる特化か，RAGによる特化か，どちらが最適なのかについて，実際に試してみないと分からない部分があります．そのため，クラウド等を利用して実際にシステムに組み込んで，お客さまにPoC（Proof of Concept）的にご利用いただく中で検証を行っていきます．

「tsuzumi」の発表から5カ月の短期間で事業展開

「tsuzumi」発表以降約5カ月という短期間で商用発表がなされましたが，具体的にはどのような事業展開がなされるのでしょうか．

NTT Com社内のGenerative AIタスクフォースに私たちのプロジェクトメンバも参画し，前述のようなプロセスにより，特化型LLMをパブリッククラウド基盤，プライベートクラウド基盤上

に構築し，お客さま個社別のソリューションとして，導入から運用までをそのサポートとともに提供していきます（**図2**）．

提供予定のソリューションについては，まずは「CX（Customer eXperience）ソリューション」「EX（employee eXperience）ソリューション」「CRX（事業継続性強化）ソリューション」の3パターンです．

具体的に「CXソリューション」では，チャットボットに加え，アバターを利用することで店頭・店舗コミュニケーションにおいて新たな顧客体験を提供する「カスタマフロントソリューション」，応対記録から必要な情報を自動抽出・要約を行うことでオペレータ業務の効率化を支援するとともに，通話内容を基にナレッジの抽出と会話サンプルの生成を行い，研修やFAQの高度化に活用することでバックヤード業務の時間削減やナレッジの高度化を支援する「コンタクトセンターソリューション」を提供します．

「EXソリューション」としては，金融・医療・行政・小売・運輸などの業界を中心に，お客さまの業界，業務に合わせたプライベート環境に生成AIの動作環境を構築することで，社内に閉じた業務マニュアルや製品仕様書，設計書など秘匿性の高いデータを学習させ，お客さまの業務プロセスに沿った業務改善に貢献し，従業員の生産性向上につながるソリューションを提供します．

「CRXソリューション」としては，ITシステム運用の自動化ソリューションに加え，お客さまのシステム情報とセキュリティ情報を学習したAIが対応アドバイスを生成することでお客さま環境に基づいたサポートを提供します．これにより，セキュリティ運用の負担を低減することが可能となるばかりではなく，マルウェア対策など年々増え続けるサイバー攻撃へのセキュリティ対応稼働の増大抑止にも対応します．

技術のコアな部分の基本スキルと情報発信が大切

技術者としてスキルの維持，スキルアップはどうしていますか．

私は2009年にNTT東日本に入社し，5年間ほどNTTネットワークサービスシステム研究所（当時）と連携してNGN（Next Generation Network）のSIP（Session Initiation Protocol）サーバの開発を行ってきました．業務では，ソフトウェアのコーディングではなく外部条件設計や検証計画等，どちらかというとドキュメントベースの工程でした．そして，NTT Comに転籍後は先端IPアーキテクチャセンタ（当時）で，ソフトウェアエンジニアとして，Webブラウザ上で音声や映像など大容量のデジタルデータをリアルタイムに送受信できる技術であるWebRTC（Web Real-Time Communication）に関するプログラミングやシステム設計を行ってきました．

NGNは通信キャリアのネットワークの世界ですが，WebRTCはインターネットの世界であり，ICE（Interactive Connectivity Establishment）といった電話の世界ではほぼ使われないプロトコルについて，IETF（Internet Engineering Task Force）のRFC（仕様書に近いドキュメント）を読みながら理解しました．さらに，プロダクト開発としてクラウド技術を深く理解したうえで，サーバサイドとフロントエンドのプログラミングも行うために，いわゆるフルスタック的なスキルも必要となり，その勉強もしました．また，アジャイルな開発スタイルでプロダクトを開発していたので，それも勉強しました．これらが今の業務のスキルとして非常に役立っています．

その後，ChatGPTが登場して世の中がさらに変わると思い，2023年7月にGenerative AIプロジェクトを正式に立ち上げましたが，短期間でチーム全体のスキルアップを図る必要があり，自分自身も含めて勉強しながら業務を進めるような状態です．しか

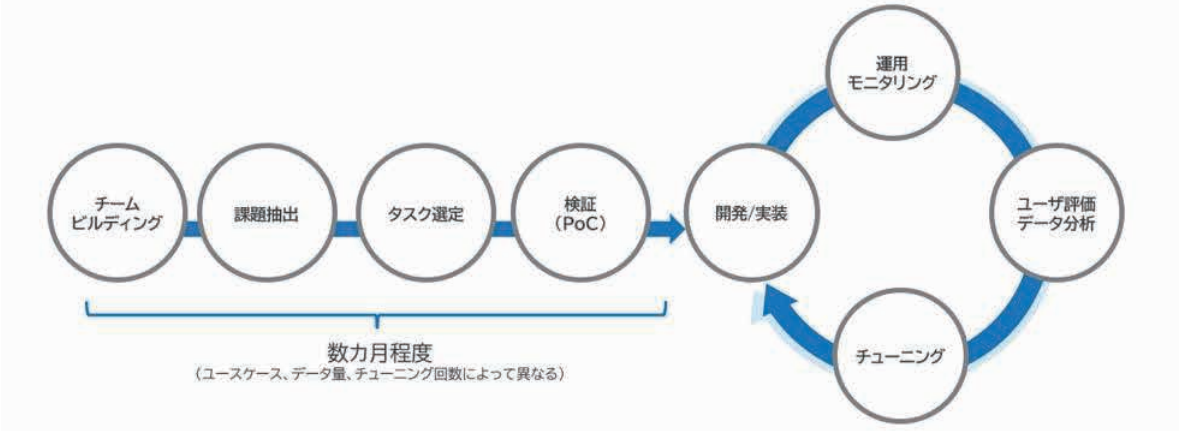

図2　「tsuzumi」等の生成AIの導入から運用までのサポートプロセス

も技術の変遷が激しい分野でもあるので，そのフォローも大変です．そこに，スキルそのものだけではなく，スキル獲得のためのプロセス等，過去の勉強・経験が役に立っています．自身が得た知見は，チーム全体に積極的に展開しています（チームメンバどうしもかなり積極的に知見を共有しています）．

開発において大切にしていることは何でしょうか．

コンピュータのOSのカーネルやネットワークアーキテクチャ，プロトコルといった技術のコアの部分に関する基本をできるだけ理解することを心掛けています．例えばリレーショナルデータベース（RDB）の場合，レコードの数が多くなると，条件付きレコード抽出の動作が遅くなります．その場合，もしRDBの基本的な仕組みを理解していると，遅くなるメカニズムがすぐに分かり，インデックス付与方法を工夫することで，それがある程度改善するとすぐ分かります．同様にソフトウェアのアルゴリズムやデータ構造等も理解できていると，新機能実装の際も非常にスムーズにできるようになります．

それから，コミュニティやカンファレンス等の場における情報発信を積極的に行っています．これにより，トップエンジニアやエキスパートが発信された情報を目にする機会が増え，それがきっかけとなってこういった方々とのコネクションが増え，気軽に意見交換ができるようになります．それにより，NTTグループに閉じない最新の情報や正しい知識が数多く入るようになります．面白いもので，自分の発信量が多いほど，情報が多く入ってきます．そして，それが例えばOSS（Open Source Software）やライブラリ採用にあたっての精度の高い口コミ的な情報となり，また，技術や施策の評価等においても，NTTグループにとらわれず非常に客観的な視座を築くことができます．

情報発信による仲間づくりと，基本スキルによりブレない技術者をめざす

将来的に何をめざして開発を続けるのでしょうか．

私自身はエンジニアでいながら，少しでも良いアウトプットを出すことができるようなプロジェクトづくり，プロジェクト連携の仕組みの実現に取り組んでいきたいと思います．私たちのアウトプットは事業部への技術支援や，新しいビジネスにつながる技術開発です．プロジェクトとして，より多くのより良いアウトプットを出していくのは，誰もが考えることではないかと思いますが，それを阻害する要因はプロセス上のボトルネックであったり，新しい取り組みであるが故に周囲の理解が追い付かないことであったりします．ボトルネック解消はプロジェクトや組織マネジメントを的確に行うことで，ある程度実現できるのではないかと思いますが，特に理解促進については，仲間や賛同者を増やすことが一番の近道ではないかと考えており，それは単なるプロジェクトマネジメントでは対応できないことです．そのためにプロジェクトそのもの，そして効果的なプロジェクト連携の仕組みをつくることが大切であり，そこに取り組んでみたいと思っています．そのうえで，チームのエンジニアやリサーチャーが，モチベーションを高め，さらにスキルを伸ばし，そして楽しいと思って技術開発に取り組んでもらえれば，良い結果につながるのではないかと思います．

社内外の技術者，パートナーへのメッセージをお願いします．

技術者の皆さん．開発を進めていくにあたり，情報収集をする機会が多いと思いますが，私の経験として，情報発信するところには情報が集まってきます．その結果，単なる情報収集ではなく，情報交換・意見交換といったところまでつながりが広がってきます．最近の開発は，単独ではなくコラボレーションにより行われることが多くなっていますが，情報交換・意見交換をした人脈やその内容等が大きく寄与することは，想像に難くないと思います．また，前述のとおり，コンピュータのOSのカーネルやアルゴリズム，ネットワークアーキテクチャ，プロトコルといった技術のコアとなっている基本を勉強して身につけておくといいと思います．これらをベースとして，その上にさまざまな技術が展開されることでサービスや商品が出来上がっているのですが，こういった技術は時代や市場とともに大きく変化していきます．ところが，基本的な部分は外的な要因により短期間で大きく変わるものはほとんどありません．したがって，この部分の技術を身につけておくことで，新しい技術の理解も早まり，技術の変遷に対しても，いい意味でブレることのない技術者になれるのではないかと思います．

パートナーの皆さん．「tsuzumi」の商用開始の発表のときに，「モデルパートナー」「ソリューションパートナー」「インテグレーションパートナー」からなるパートナーシッププログラムの募集を紹介させていただきました．私たちのチームでは，パートナー企業が保有している業界特化データと「tsuzumi」を組み合わせることで，業界・業務に特化した新しいLLMを共同で構築し，展開する「モデルパートナー」を募集しています．業界・業務特化型の新しいLLMの構築にご一緒していただけるよう，ご応募をお待ちしています．